A comparison of spatial scan methods for cluster detection

Joshua P. French[1], Mohammad Meysami[2], Lauren M. Hall[3], Nicholas E. Weaver[1], Minh C. Nguyen[4], and Lee Panter

[1]Department of Mathematical and Statistical Sciences, University of Colorado Denver, Colorado, USA
[2]Department of Mathematics, Clarkson University, New York, USA
[3]Neptune and Company, Colorado, USA
[4]Metropolitan State University, Colorado, USA

Abstract

Spatial scan methods are extremely popular for identifying disease clusters using disease count data. The original circular scan method proposed by Kulldorff [1] is simple to implement, is computationally inexpensive to apply, and has high power for detecting circular clusters; however, it can struggle to identify non-circular clusters. Many extensions of the original method have been proposed to better detect irregularly-shaped clusters. We briefly describe several popular spatial scan method extensions (e.g., Upper Level Set, Flexibly-shaped, Dynamic Minimum Spanning Tree, Fast Subset, etc.). We then compare the performance of the various methods using power, sensitivity, positive predictive value, and overall accuracy by applying these methods to 126 publicly-available benchmark data sets based on 46 different cluster shapes. The comparisons go into more depth and include more methods than any previous studies of this topic; many of the methods have never been directly compared. The comprehensiveness of our study allows us to draw reliable conclusions and make concrete recommendations about the best performing methods. R packages and scripts are provided to make results reproducible.

Keywords: spatial scan statistic, disease cluster identification, most likely cluster, likelihood ratio statistics, power

1 Introduction

Data related to health, crime, and other events of interest are frequently reported as counts within pre-specified enumeration regions such as counties or census tracts. This helps preserve the privacy of individuals associated with the event of interest while allowing for pattern detection. Researchers are often interested in identifying clusters, which are collections of (usually) contiguous regions where incidence rates are higher than those of surrounding regions.

Statistical methods for cluster detection are frequently proposed in the context of disease outbreak. Waller and Gotway [2], Tango [3], and Bivand et al. [4] provide helpful overviews of many of the popular methods available for cluster detection. Some early, well-known methods for cluster detection include Moran's I [5], Geary's C [6], and their subsequent extensions [7, 8, 9]. Exploratory methods for cluster detection that do not make statistical inference include the Geographical Analysis Machine[10] and a method based on overlapping local incidence proportions [11]. Turnbull et al. [12] and Besag and Newell [13] proposed statistical tests for cluster identification, each with a different approach for accounting for the variability in incidence rates across regions. Tango [3] and Mclafferty [14] provide many examples of methodological development in disease cluster identification since that time. In what follows, we focus our discussion toward a specific stream of research in disease cluster detection, which are broadly known as spatial scan methods.

Two main weaknesses prevalent with early cluster detection techniques were that they (i) were global tests that identified a general discrepancy between observed and expected incidence rates, but failed to identify a specific set of regions having an unusually high incidence rate or (ii) did not satisfactorily address the problem of multiple comparisons. Kulldorff and Nagarwalla [15] proposed the spatial scan method, which addressed both of these issues in a well-defined statistical framework. The utility of this method was quickly recognized,

and the method became a popular choice for disease cluster identification. As of early December 2021, the Kulldorff and Nagarwalla [15] article had over 1,700 citations. Kulldorff [1] provided additional exposition of the spatial scan method and had over 4,100 citations [16] as of December 2021. Despite the fact that the spatial scan method is over 20 years old, it continues to be widely applied even today.

The spatial scan method's popularity stems from its simplicity, computational efficiency, the availability of a free implementation of the methodology in SatScan [17], and its power to detect disease clusters. Naturally, the popularity of the spatial scan method encouraged other researchers to propose extensions that addressed different contexts or improved the accuracy of the method. These "scan methods" all seek to identify the largest value of a likelihood ratio statistic by scanning over a set of candidate zones, but differ in their choice of candidate zones.

We aim to provide a detailed comparison of numerous scan methods in this article [18, 19, 20, 21, 22, 23]. In this research, we make new and objective comparisons of the different well-known spatial scan methods and provide specific recommendations on their performance. The substantial novelty of this research involves comparing more scan methods than has ever been done before as well as utilizing a more extensive set of benchmark data. Many of the methods have never had their performance directly compared, and previous comparisons were based on smaller subsets of benchmark data sets generated by the method creators. With the purpose of providing a fair, objective comparison, we generated 45 additional data sets that were not predisposed to favor any of the methods described in this paper. Consequently, the evaluation and conclusions we present are meant to be more comprehensive and unbiased. While some of the results and implementations are available through previous research, no integrated software has been developed to allow researchers to compare methodologies in a simple manner. We have created the **smerc** R package [24] to provide a free, open-source implementation of the various benchmarked methods. This package enables researchers to reproduce the results and outputs of each spatial scan method. Making our research reproducible strengthens the veracity of our results, and will also encourage continued investigation into new methods for cluster detection.

The structure of this paper is as follows. In Section 2, we describe the original spatial scan method in further detail. We then describe many of the subsequent extensions proposed for identifying disease clusters using regional count data. In Section 3, we summarize the results when applying the scan methods to 81 benchmark data sets made available by Kulldorff et al. [25] and Duzmal et al. [26], as well as the 45 additional benchmark data specifically generated for this study. In Section 4, we discuss our findings, making specific recommendations based on our comprehensive analysis. In the Supplementary Information Section, the complete benchmark results are available in a series of tables and graphics.

2 Methodology

Consider a study area A that is partitioned into N disjoint regions $R_1, R_2, \ldots, R_N$, with $A = \{R_1, R_2, \ldots, R_N\}$. The at-risk populations of the regions are denoted $n_1, n_2, \ldots, n_N$, respectively, with $n_A = \sum_{i=1}^{N} n_i$ denoting the total population across the study area. The observed number of disease cases for the regions are denoted $y_1, y_2, \ldots, y_N$, with $y_A = \sum_{i=1}^{N} y_i$ denoting the total case count across the study area, and $y = (y_1, y_2, \ldots, y_N)$ denoting the vector of observed cases. Let $\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_N$ denote representative locations within regions $R_1, R_2, \ldots, R_N$, where $\mathbf{s}_i := (s_{i1}, s_{i2})$ is the vector of coordinates in spatial dimensions 1 and 2, respectively. We refer to $\mathbf{s}_i$ as the centroid of region R_i.

Case counts are most commonly modeled using a Binomial or Poisson distribution. Let θ_i represent disease risk in region R_i which refers to the risk of an individual being affected in region R_i. If the case counts are modeled by a Binomial random variable, then

$$y_i \overset{\text{indep.}}{\sim} \text{Binomial}(n_i, \theta_i), \quad i = 1, 2, \ldots, N. \tag{1}$$

If the case counts are modeled by a Poisson random variable, then

$$y_i \overset{\text{indep.}}{\sim} \text{Poisson}(n_i \theta_i), \quad i = 1, 2, \ldots, N. \tag{2}$$

We intend to test whether a subset $Z \subset A$ has elevated risk compared to regions in Z^c, where Z^c is the complement of Z. Let θ_Z denote the risk of disease for regions in Z and θ_0 denote the risk of disease for regions

in Z^c. Formally, we wish to test $H_0 : \theta_Z = \theta_0$ against $H_a : \theta_Z > \theta_0$. Define

$$y_Z = \sum_{i:R_i \in Z} y_i \quad \text{and} \quad n_Z = \sum_{i:R_i \in Z} n_i.$$

Assuming the case counts have the Binomial distribution in Equation (1), Kulldorff and Nagarwalla [15] derived a likelihood ratio statistic as

$$T_Z^B = \frac{\left(\frac{y_Z}{n_Z}\right)^{y_Z} \left(\frac{n_Z - y_Z}{n_Z}\right)^{n_Z - y_Z} \left(\frac{y_A - y_Z}{n_A - n_Z}\right)^{y_A - y_Z} \left(\frac{n_A - n_Z - (y_A - y_Z)}{n_A - n_Z}\right)^{n_A - n_Z - (y_A - y_Z)}}{\frac{(y_A)^{y_A}(n_A - y_A)^{n_A - y_A}}{(n_A)^{n_A}}} I\left(\frac{y_Z}{n_Z} > \frac{y_A - y_Z}{n_A - n_Z}\right), \tag{3}$$

and 1 otherwise, where $I(\cdot)$ is the indicator function (cf., Duczmal and Assunção [27]). Alternatively, assuming the case counts have the Poisson distribution in Equation (2), Kulldorff [1] derived a likelihood ratio statistic as

$$T_Z^P = \frac{\left(\frac{y_Z}{n_Z}\right)^{y_Z} \left(\frac{y_A - y_Z}{n_A - n_Z}\right)^{y_A - y_Z}}{\left(\frac{y_A}{n_A}\right)^{y_A}} I\left(\frac{y_Z}{n_Z} > \frac{y_A - y_Z}{n_A - n_Z}\right), \tag{4}$$

and 1 otherwise. We note that an alternate form of Equation (4) based on expected counts (cf., Waller and Gotway [2] and Tango and Takahashi [21]) is frequently used because it allows the researcher to estimate the expected value of each region using relevant covariate information. Kulldorff [1] proposed that the scan test be implemented using the test statistic

$$T = \sup_{Z \in \mathcal{Z}} T_Z^D, \tag{5}$$

where T_Z^D is the likelihood ratio statistic of zone Z derived from the assumed response distribution (typically T_Z^B or T_Z^P) and $\mathcal{Z}$ is the set of candidate zones under consideration. The candidate zone associated with the largest likelihood ratio statistic is referred to as the *most likely cluster*, and we denote the most likely cluster by Z_{mlc}.

Monte Carlo methods are typically used to assess the significance of the most likely cluster. B data sets are simulated independently under the null hypothesis that $\theta_i = \theta_0$ for $i \in \{1, 2, \ldots, N\}$. Typically, the number of cases in each simulated data set is fixed at y_A. Thus, the ith simulated set of cases, $\tilde{y}^i$, has a Multinomial distribution with y_A trials and the probability of a trial being assigned to region R_i is n_i/n_A, $i = 1, 2, \ldots, N$. For each simulated data set $\tilde{y}^i$, Equation (5) is used to determine the largest test statistic, which we denote T_i. The p-value for the test is computed as

$$p = \frac{1 + \sum_{i=1}^{B} I(T_i > T)}{B + 1}. \tag{6}$$

If the p-value associated with T is less than the significance level α, then we believe that Z_{mlc} is an actual cluster of regions with higher risk than regions outside the cluster. If T_Z refers to the test statistic associated with a particular candidate zone Z, then secondary clusters are identified by replacing T in Equation (6) with T_Z, computing the associated p-value, and determining the candidate zones with sufficiently small p-values.

The substantive difference between different scan methods is the approach for constructing $\mathcal{Z}$, the set of candidate zones under consideration. For even moderately large N, determining all possible combinations of zones (typically connected) is computationally infeasible. Thus, researchers have proposed many different approaches for strategically choosing candidate zones to make computation feasible, but flexible enough to identify clusters of many different shapes.

We next outline the basic details of the scan methods we will compare. In general, each method will sequentially increase the size of candidate zones by starting with a single region as a candidate zone and then augmenting that candidate zone with new regions until some constraint is violated. The most common constraint is that no more than 50% of the total population may be in a candidate zone. Other common constraints are that the candidate zones are subsets of the k nearest neighbors of each starting centroid or that the intercentroid distance between any two regions in the candidate zone is no more than some maximum

geographic size. Regardless, we let k_i denote the maximum number of regions that can be added to a candidate zone starting with region R_i (in the context of a specific scan method). For simplicity, we assume a population constraint is being used, but the principles apply to any similar constraint the researcher chooses to use. In general, we will let I refer to the index of a region or centroid. In order to prevent notation for indices from becoming overly complex, the notation for indices is slightly redefined between subsections, i.e., the specific notation for indices in one subsection is likely not the same as the notation in the next section.

2.1 The circular scan test

Kulldorff and Nagarwalla [15] and Kulldorff [1] proposed the original spatial scan test. It is frequently referred to as the *circular* scan test, as the candidate zones in $\mathcal{Z}$ are circular in shape. The candidate zones used in the circular scan test are determined by sequentially adding the nearest neighbors of each starting region to construct new candidate zones, with the constraint that no more than 50% of the total population is in a candidate zone.

We briefly describe the set of candidate zones for the circular scan method in more detail. Let $I_{i(j)}$ denote the index of the jth nearest region to region R_i in terms of intercentroid distance. By definition, $I_{i(1)} = i$, since the starting region has a distance of zero with itself. For each starting centroid $\mathbf{s}_i$, $i = 1,2,\ldots,N$, we construct the sequence of candidate zones $\{R_{I_{i(1)}}\},\{R_{I_{i(1)}},R_{I_{i(2)}}\},\ldots,\{R_{I_{i(1)}},R_{I_{i(2)}},\cdots,R_{I_{i(k_i)}}\}$, where $I_{i(k_i)}$ denotes the final index for which the population constraint is satisfied by the sequence of candidate zones extending from region R_i. The complete set of candidate zones for the circular scan method is

$$\mathcal{Z} = \bigcup_{i=1}^{N}\bigcup_{j=1}^{k_i}\{R_{I_{i(1)}},R_{I_{i(2)}},\ldots,R_{I_{i(j)}}\}.$$

The circular scan test is still an extremely popular method for cluster detection. It is relatively simple to understand, is very fast to implement due to the special structure of the candidate zones, and is available as part of the free, publicly-available SaTScan software [17].

2.2 The elliptic scan test

Kulldorff et al. [25] showed that the circular spatial scan method has high power to detect circular clusters, but it may struggle in detecting irregularly-shaped clusters. Kulldorff et al. [18] proposed supplementing the circular candidate zones of the circular scan method with elliptic candidate zones. While the circular scan method is direction free, the regions included in an elliptic candidate zone may vary considerably if the ellipse is rotated around its center.

An ellipse can be characterized by three parameters: its origin, $\mathbf{s}$, its shape, ς, and its angle, ϕ. The origin is the center of the ellipse. The shape of the ellipse is defined as the ratio between the lengths of its major axis, b, and its minor axis, a, i.e., $\varsigma = b/a$. The smallest value of ς is 1, which represents a circular shape. Increasing the value of ς results in a narrower ellipse. The angle ϕ is the angle between the major axis and the horizontal axis. Figure 1 depicts a representative ellipse.

We now describe how candidate zones are constructed using an ellipse. Let $\mathbf{s}_i$ be the origin of the ellipse, and fix the shape and angle parameters, ς and ϕ. Note that this implies the major axis $b = a/\varsigma$. The ellipse is enlarged by increasing the lengths of its major and minor axes while maintaining the shape ς. A region R_j is "included" in the ellipse if its centroid $\mathbf{s}_j$ lies within the ellipse, i.e., if it satisfies the constraint

$$\frac{[\cos(\phi)(\mathbf{s}_{j1}-\mathbf{s}_{i1})+\sin(\phi)(\mathbf{s}_{j2}-\mathbf{s}_{i2})]^2}{a^2}+\frac{[\cos(\theta)(\mathbf{s}_{j1}-\mathbf{s}_{i1})-\sin(\phi)(\mathbf{s}_{j2}-\mathbf{s}_{i2})]^2}{(a/\varsigma)^2}\leq 1. \tag{7}$$

The ellipse is increased in size until the combined population of all regions with centroids inside the ellipse exceeds 50% of the total population. Let $a_{ij}^{\varsigma,\phi}$ denote the minimum length of the minor axis needed for $\mathbf{s}_j$ to satisfy the constraints of Equation (7) when the ellipse originates from centroid $\mathbf{s}_i$. Additionally, let $a_{i(j)}^{\varsigma,\phi}$ denote the jth largest length $a_{ij}^{\varsigma,\phi}$ and $I_{i(j)}^{\varsigma,\phi}$ denote the index of the region associated with $a_{i(j)}^{\varsigma,\phi}$. We define the sequence of candidate zones starting from centroid $\mathbf{s}_i$ with shape ς and angle ϕ by

$$\mathcal{Z}(\mathbf{s}_i,\varsigma,\phi)=\{\{R_{I_{i(1)}^{\varsigma,\phi}}\},\{R_{I_{i(1)}^{\varsigma,\phi}}\cup R_{I_{i(2)}^{\varsigma,\phi}}\},\ldots,\{R_{I_{i(1)}^{\varsigma,\phi}}\cup R_{I_{i(2)}^{\varsigma,\phi}}\cup\cdots\cup R_{I_{i(k_i)}^{\varsigma,\phi}}\}\},$$

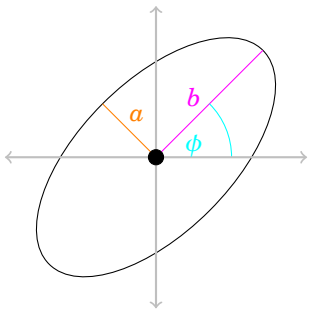

Figure 1: An ellipse centered at centroid **s**, rotated with angle ϕ, with minor axis a and major axis b. The minor axis a is shown in orange, the major axis b is shown in magenta, the rotation angle ϕ is shown in cyan, and the origin **s** is indicated by the black dot.

where k_i denotes the last index such that the population constraints are satisfied. Note that we intentionally suppress the dependence of k_i on ς and ϕ for simplicity, as k_i may differ for each combination of ς and ϕ, and multiple combinations are generally considered. Kulldorff et al. [18] recommended that the set of candidate zones in $\mathcal{Z}$ be comprised of the zones obtained from considering many combinations of ς and ϕ and that for a fixed ς, a sequence of angles should be chosen so that at least 70% of each ellipse overlaps the ellipse associated with the next angle. Let Σ denote the set of shapes to consider and Φ_ς denote the set of angles considered for shape ς. The set of zones considered by the elliptic scan test is

$$\mathcal{Z} = \bigcup_{i=1}^{N} \bigcup_{\varsigma\in\Sigma} \bigcup_{\phi\in\Phi_\varsigma} \mathcal{Z}(\mathbf{s}_i, \varsigma, \phi).$$

When the shape is large, the ellipse is long and narrow, and regions in a zone may not be neighboring. To prevent this eccentric tendency, Kulldorff etl al. [18] recommended a penalized version of the Poisson test statistic in Equation (4):

$$T_{adj} = \sup_{Z\in\mathcal{Z}} \left\{ T_Z^P \left(\frac{4\varsigma}{(\varsigma+1)^2} \right)^{\lambda} \right\},$$

where $\lambda \geq 0$ is a penalty parameter. In the circular case, when $\varsigma = 1$, T_{adj} is the same as the original Poisson test statistic since there is no need to fix the eccentricity. Otherwise, $(\varsigma+1)^2 \geq 4\varsigma$ for all values of $\varsigma \geq 1$, thus the penalty is stronger when increasing λ with a fixed $\varsigma > 1$. Moreover, when $\lambda = 0$, no penalty is applied to the test statistic, and when $\lambda \to \infty$, the penalty is so strong that only circular clusters are considered [18].

2.3 The upper level set (ULS) scan test

The circular and elliptic scan methods impose an artificial restriction on the shape of zones evaluated for increased risk. Each zone under consideration must conform to the geometric restrictions of the underlying search methodology, so for the circular and elliptic scan methods, the clusters discovered would be either circular or elliptical, respectively.

Patil and Taillie[19] proposed the ULS scan method for detecting arbitrarily-shaped clusters based on the underlying spatial connectivity of the study area A, combined with a population-scaled response value to determine a reduced parameter space over which an exhaustive search can be completed. We define the population-scaled response values as

$$G_i = y_i/n_i, \quad i = 1,2,\ldots,N,$$

with $G := \{G_1, G_2, \ldots, G_N\}$. We define the distinct values of G to be the values $r_1 > r_2 > \cdots > r_M$, where $M \leq N$. Additionally, for each value of r_j we define index sets

$$I_j = \{i \in \{1,2,\ldots,N\} : G_i = r_j\}, \quad j = 1,\ldots,M.$$

From each I_j, we may define an Upper Level Set U_j for $j = 1,\ldots,M$ by

$$\begin{aligned} U_j &= I_1 \cup I_2 \cup \cdots \cup I_j \\ &= \{i \in 1,2,\ldots,N : G_i \geq r_j\}. \end{aligned}$$

The Upper Level Set U_j corresponding to the distinct population-scaled response value r_j can be interpreted as the set of regional indices that have a population-scaled response value at least as large as r_j. This interpretation means that for any given $j = 1,\ldots,M$, the set U_j contains only those regions with the highest values of population-scaled response values.

Special consideration is given to each of the contiguous subsets within U_j. Let

$$C_j = \{C_{j(1)}, C_{j(2)}, \ldots, C_{j(k_j)}\}, \quad j = 1,2,\ldots,M,$$

be the set of connected components of U_j, where a "connected component" $C_{j(l)}$ is simply a contiguous set of regions in U_j. The components are constructed to be as large as possible, but with no overlap between any two components, i.e., $C_{j(l_1)} \cap C_{j(l_2)} = \emptyset$ when $l_1 \neq l_2$. According to this definition, we know that $1 \leq k_j \leq |U_j|$, where $|U_j|$ is the number of regions in U_j. The defining properties of U_j imply that each connected component $C_{j(l)}$ has elevated population-scaled response values in comparison to those regions not included in U_j. This property, along with the contiguity of each $C_{j(l)}$, qualifies the connected components as candidates for the most likely cluster. We therefore define the new reduced parameter space for the ULS test as

$$\mathcal{Z} = \bigcup_{j=1}^{M} \bigcup_{l=1}^{k_j} C_{j(l)}.$$

Patil and Taillie [19] and Patil et al. [28] provided helpful examples and graphics to aid in understanding the structure of the candidate zones.

2.4 The flexibly-shaped scan test

Tango and Takahashi [20] proposed the flexibly-shaped scan method, which allows for non-circular clusters by searching among all connected subsets of regions under certain constraints.

Let $\mathcal{Z}_C$ and $\mathcal{Z}_F$ denote the set of all zones that need to be scanned by the circular and flexibly-shaped scan methods, respectively. Let $\mathcal{N}_k(R_i)$ denote the k nearest neighbors of region R_i, including the region itself. A k-nearest neighbors version of $\mathcal{Z}_C$ is obtained by determining the union of the zones created when sequentially adding each of the k nearest neighbors (in order of distance) to each starting region R_i. This union results in a maximum of Nk candidate zones in $\mathcal{Z}_C$. In addition to the circular candidate zones in $\mathcal{Z}_C$, the flexibly-shaped scan method also considers any candidate zone $Z^* \subseteq \mathcal{N}_k(R_i)$ in which there is a connected path between all regions in Z^*. Consequently, $\mathcal{Z}_C \subset \mathcal{Z}_F$ and $\mathcal{Z}_F$ has substantially more candidate zones than $\mathcal{Z}_C$. Tango and Takahashi [20] described a computationally efficient approach for constructing $\mathcal{Z}_F$.

The flexibly-shaped scan method considers all circular and irregularly-shaped zones up to a maximum neighborhood size, allowing the flexibly-shaped method to detect non-circular clusters. However, the computation time need to construct $\mathcal{Z}_F$ increases exponentially as k gets larger. This may require a researcher to use a relatively small value of k in order to apply the method in a reasonable amount of time, limiting the method's ability to detect larger clusters. Moreover, in situations where the true cluster is circular, the flexibly-shaped method tends to detect clusters larger than the true cluster.

2.5 The restricted flexible scan test

Tango [29] proposed a restricted version of the flexibly-shaped scan method. The restricted method was further studied by Tango and Takahashi [21]. This restricted method is more computationally friendly than the original flexibly-shaped scan method while simultaneously allowing for larger clusters to be detected. Tango [29] also noted that the flexibly-shaped scan method tends to detect much larger clusters than the true cluster because the method allows for candidate zones to include regions with non-elevated risk. The restricted method eliminates these regions from the candidate zones.

Tango [29] and Tango and Takahashi [21] adjusted Equations (4) and (5) to create the restricted likelihood spatial scan statistic,

$$\sup_{Z \in \mathcal{Z}} \left\{ T_Z^P \prod_{i:R_i \in Z} I(p_i < \alpha_1) \right\}, \tag{8}$$

where p_i is the middle p-value given by

$$p_i = P(Y_i \geq y_i + 1) + \frac{1}{2} P(Y_i = y_i), \tag{9}$$

α_1 is a pre-specified significance level, and $Y_i \sim \text{Poisson}(e_i)$, where e_i is the expected number of cases in region R_i under the null hypothesis. Based on Equations (8) and (9), the restricted flexible scan method only considers candidate zones where all included regions have elevated risk. Otherwise, $I(p_i < \alpha_1)$ becomes zero and the entire candidate zone is insignificant. Excluding candidate zones from consideration that have regions with non-elevated risk reduces the search space of the method, making the computational load lighter than the original method, while still retaining much of its flexibility. One down side of the restricted version of the test is that the candidate zones are data-driven, meaning that the candidate zones must be determined for the original data set and each data set simulated under the null hypothesis. Additionally, this method tends to "disconnect" different areas of a cluster if an important connecting region has large p_i, making it more difficult to detect large clusters.

2.6 Minimum spanning trees

To address the limitations of the circular scan method for detecting non-circular clusters, another set of techniques for identifying irregularly shaped clusters are classified as minimum spanning trees and introduced by Assunção et al. [22]. Their goal was to efficiently scan a subset of candidate zones without needing an arbitrarily selected tuning parameter (as seen in other methods). We briefly describe the intuition for the approach as well as the construction of candidate zones.

The key focus of this method is determining the candidate zones that should be included within the subset to be scanned. Assunção et al. [22] suggested representing the data as a weighted graph, $\mathcal{G} = (\mathbf{V}, \mathbf{E}, \mathbf{W})$, where $\mathbf{V}$ is the set of region centroids, $\mathbf{E}$ is the set of edges linking regions that share a border, and $\mathbf{W}$ is the set of edge weights. Since the sets $\mathbf{V}$ and $\mathbf{E}$ are fixed, the goal is to determine edge weights so that they provide information about likely clusters. Without loss of generality, we specify large weights to indicate regions that should be clustered together (i.e., they have similar disease risk) and small weights to indicate regions that should not be clustered together (i.e., their disease rates are dissimilar). A good candidate zone for the true cluster can be represented by connected regions with a large overall weight. Therefore, constructing a minimum spanning tree of $\mathcal{G}$ identifies a small subset of zones to be considered for analysis.

Assunção et al. [22] developed two methods to accomplish the above goal. Costa et al. [30] proposed three modifications to the construction of minimum spanning trees in an attempt to resolve weaknesses in the previous methods. We introduce one method from Assunção et al. [22] and the three methods from Costa et al. [30] in the following subsections.

2.6.1 The dynamic minimum spanning tree scan test

Assunção et al. [22] developed the dynamic minimum spanning tree (dMST) method as an initial implementation of minimum spanning trees for cluster detection. The dMST method employs a greedy algorithm to construct candidate zones by augmenting a current candidate zone with the connected region that most increases the likelihood ratio scan statistic. The algorithm for constructing the candidate zones of the dMST scan method is as follows:

1. Select an initial region R_i as a candidate zone Z.
2. Compute the statistic T_Z^D for the candidate zone.
3. Identify all regions connected with the current candidate zone Z. Let $\mathcal{N}_Z$ be the set of regions connected with Z.

4. For each $R_j \in \mathcal{N}_Z$:

   (a) Let $Z_j^{'} = R_j \cup Z$.

   (b) Compute the statistic $T^D_{Z_j^{'}}$.

5. Based on the previous step, update the candidate zone as $Z = \text{argmax}_{Z_j^{'}} T^D_{Z_j^{'}}$.

6. Repeat steps 3, 4, and 5 until a stopping criterion is met.

7. Repeat steps 1-6 for $i = 1, 2, \ldots, N$.

The most likely cluster is the zone Z with the largest T^D_Z computed in the algorithm above.

Assunção et al. [22] noted that the dMST scan method tends to favor large, sprawling clusters, which they call the “octopus effect". The dMST scan method typically has longer execution times than the methods previously discussed, as candidate zones must be uniquely determined for the observed data set and each simulated data set.

2.6.2 The early stopping dynamic minimum spanning tree scan test

Costa et al. [30] proposed the early stopping dynamic minimum spanning tree (edMST) scan test to improve on the original dMST method by adding an early-stopping criterion. In step 4 of the algorithm in Section 2.6.1, a new region will be added to the current candidate zone only if the addition increases T^D_Z. This effectively stops the algorithm from branching off into long-reaching clusters, a common issue with the dMST scan test. As a consequence of this modification, the cluster of interest for each starting region is simply the candidate zone that is present when the algorithm terminates. This efficiently reduces the computational time and complexity of the dMST test.

2.6.3 The double connection scan test

Costa et al. [30] also proposed a “double connection" (DC) scan method. Similar to the edMST scan method, the DC scan test includes the same early-stopping criterion mentioned in Section 2.6.2. To increase the compactness of the candidate zones, the DC scan test adds an additional restriction: when constructing the candidate zones, new connecting regions will be considered for addition only if they share a border with at least two regions in the current candidate zone (or a single region if the candidate zone consists of only a single region). The belief is that this additional constraint will force the candidate zones to maintain a more plausible shape throughout construction. As before, the early stopping criterion significantly reduces the computational time required to complete the algorithm. Additionally, the DC restriction reduces the number of candidate regions considered at each step of the algorithm

2.6.4 The maximum linkage scan test

Costa et al. [30] also proposed the maximum linkage (m-link) method in an attempt to remove the octopus effect without using an early stopping criterion. Suppose we are constructing the set of candidate zones as described for the dMST method in Section 2.6.1. The m-link scan test will seek to add a region not currently in the candidate zone only if it shares the maximum number of connections (i.e., borders) with the candidate zone among all regions bordering the current candidate zone. If two or more regions share the maximum number of connections, then the region that most increases T^D_Z is selected. A stopping criterion is required (typically a population upperbound or maximum number of regions, but not the early stopping criterion used by the edMST and DC scan tests) to determine when the algorithm for constructing candidate zones should be terminated. This method is similar in computational complexity to the dMST method as a similar number of calculations are made in both methods.

2.7 The fast subset scan test

While methods such as the elliptic or flexibly-shaped scan tests have high power to detect disease clusters by searching a large set of candidate zones, they tend to suffer from long computation times, with the number of potential clusters increasing rapidly as the size of the study area increases. Neill [23] proposed a fast subset scan, which seeks the maximum scan statistic over a considerably smaller set of potential clusters, doing so in linear time.

In the fast subset scan, each region $R_i \in A$ is assigned a priority based on a pre-specified priority function $G(i)$, and the regions are sorted according to priority. Let $R_{(j)}$ be the region with the j^{th} highest priority. Neill defines a statistic T^D and associated priority function G to have the linear time subset scanning (LTSS) property if and only if

$$\max_{Z \subseteq \mathcal{Z}} \{T_Z^D\} = \max_{j=1,\ldots,N} \left\{ T^D_{\{R_{(1)} \cup \cdots \cup R_{(j)}\}} \right\}.$$

In other words, if the maximum over all statistics occurs on a subset of the first j regions ordered by priority, then that statistic T^D and priority function G have the LTSS property, and the global maximum of T^D can be found in linear time. Neill proves that Kulldorff's scan statistic as defined by Equation (4) has the LTSS property when accompanied by the priority function $G(i) = y_i/e_i$.

Neill [23] implements the unrestricted fast subset scan method as follows:

1. Calculate the expected counts e_i for $i = 1, 2, \ldots, N$.
2. Calculate the priority score $G(i) = y_i/e_i$ for $i = 1, 2, \ldots, N$.
3. Sort the regions in descending priority order. This can be done in $O(N \log N)$ time.
4. Calculate $T^D_{\{R_{(1)}, R_{(2)}, \ldots, R_{(j)}\}}$ for $j = 1, 2, \ldots, N$, for a total of N statistics.
5. Take the maximum over the N statistics.

As only N statistics need to be calculated, the maximum scan statistic can be computed in $O(N)$ time.

Neill's fast subset scan is similar to the ULS method in that both order the regions by priority and consider subsets based on the ordering. However, the fast subset scan does not impose any connectivity constraints. This means that the unrestricted fast scan often returns a set of disconnected regions as the "most likely cluster" due to those regions having the highest priority scores.

Neill [23] proposed two variants of the fast subset scan that impose proximity constraints, forcing the fast subset scan to return a set of spatially close regions as the most likely cluster, though the regions may still be disconnected. The fixed k neighborhood forms local neighborhoods using each R_i and its $k-1$ nearest neighbors. The fast subset scan is performed on each of the N local neighborhoods, and the maximum taken over all the statistics. The fixed r method is similar in execution, but the local neighborhoods are defined by all regions within a fixed distance r of region R_i. Both the fixed k and the fixed r local subset scans can be repeated with multiple values of k and r, which Neill names the multiscan k and multiscan r methods. Once each set of scans has been performed, a Pareto set of the scan statistics is created containing all potential clusters Z of size k_Z or r_Z, such that no other cluster Z' of size $k_{Z'}$ or Z' has the following properties:

$$T^D_{Z'} > T^D_Z \text{ and } k_{Z'} \leq k_Z \text{ or } r_{Z'} \leq r_Z,$$
$$T^D_{Z'} = T^D_Z \text{ and } k_{Z'} < k_Z \text{ or } r_{Z'} < r_Z.$$

The most likely cluster is then selected from those in the Pareto set by computing $T^D_Z - L_{k_Z}$ or $T^D_Z - L_{r_Z}$ for some constant L and taking the maximum. Larger values of L penalize larger clusters and result in smaller clusters, while values of L at or near zero penalize size less and result in larger clusters. The multiscan k and multiscan r approaches take considerably more time to compute than the unrestricted fast scan, as instead of searching N potential clusters, the multiscan k searches kN^2 potential clusters for each value of k, and the multiscan r searches $\bar{k}N^2$ potential clusters for each value of r, where $\bar{k}$ is the average neighborhood size for that r.

2.8 Other methods

Other scan-related methods have been proposed that we have not considered. We briefly mention several of them.

Duczmal and Assunção [27] used a simulated annealing algorithm to select good candidate zones. Duczmal et al. [31] used a (mono-objective) genetic algorithm to select quality candidate zones, while Duczmal et al. [32] used a multi-objective genetic algorithm to identify candidate zones. In each case, tuning parameters must be specified to implement the algorithms. Additionally, the implementation of these methods is arguably more complicated than those of the previously discussed methods.

Assunção et al. [22] also proposed a static minimum spanning tree, but it was shown to be substantially less effective than the dMST method.

Patil et al. [28] proposed a Progressive upper-level set (PULSE) method intended to improve on the ULS method [19], but the algorithm for choosing candidate sets becomes much more complicated and no implementation of the algorithm is currently available.

Murray et al. [33] proposed a method for identifying promising candidate zones in such a way that the statistic T_Z^D is maximized under certain conditions. The method is similar in spirit to the fast subset scan method of Neill [23], but makes different assumptions. The original implementation of this method required proprietary software that may not be readily available to many researchers.

There are certainly other scan-based cluster detection methods that we have not mentioned due to the sheer scope of scan-based methods. However, we have tried to highlight the ones most popular in the academic literature.

3 Benchmark evaluations

3.1 Background

We now benchmark the methods described in Section 2 using publicly-available benchmark data sets constructed by Kulldorff et al. [25] and Duczmal et al. [26]. The data sets are inspired by breast cancer mortality data from the northeastern United States during the years 1988-1992. The data were previously studied by Kulldorff et al. [34]. The original breast cancer data set included mortality counts for 245 regions spread throughout Connecticut, Delaware, Maine, Maryland, Massachusetts, New Hampshire, New Jersey, New York, Pennsylvania, Rhode Island, Vermont, and the District of Columbia. The population in each region corresponds to the number of women recorded in the 1990 United States Census. A centroid coordinate is associated with each region. At the time of this writing, the data constructed by Kulldorff et al. [25] are available at `https://www.satscan.org/datasets/nebenchmark/index.html`, while the data constructed by Duzmal et al. [26] are available at `http://www.est.ufmg.br/~duczmal/databases/`.

Kulldorff et al. [25] provided data sets based on 70 different cluster models. Kulldorff et al. [25] simulated clusters centered around three area types: a rural area (Grand Isle County in Vermont), a "mixed" area (Allegheny County in Pennsylvania), and an urban area (New York City). Clusters with 1, 2, 4, 8, and 16 regions were simulated for each area type. For each of the cluster models, the risk within the cluster was greater than the risk outside the cluster. In addition, Kulldorff et al. [25] created additional scenarios by combining the clusters previously described. Specifically, there were scenarios where rural and urban, rural and mixed, mixed and urban, and rural, mixed, and urban clusters occurred simultaneously, each occurring with the same number of regions in each cluster. For each cluster model, data sets were simulated with both 600 and 6000 cases. The cases were distributed across the regions in proportion to population, except in the clusters, which had a greater risk.

All 70 cluster models used by Kulldorff et al. [25] were circular in shape. Duczmal et al. [26] generated irregularly-shaped clusters (referred to as a, b, …, k) using 11 different models. To have a more extensive comparison and avoid biasing the results in favor of the circular scan method, we generated 45 additional irregularly-shaped clustering models. Three different sets of irregularly-shaped clustering models, irural, imixed, and iurban (i.e., **i**rregularly-shaped **urban** clustering model) are generated. Each clustering model irural, imixed, and iurban contains 2-16 regions (counties). For the irregularly-shaped clusters, the number of cases observed across the study area was fixed at 600 for each data set. For each of the 126 different cluster models, 10,000 data sets were generated. In Figure 2, the top-left panel displays the clusters simulated by

Kulldorff et al. [25] that have circular shapes. The differing shades of color for the mixed, urban, and rural regions are intended to help delineate the clusters consisting of 1, 2, 4, 8, or 16 regions. Other panels in Figure 2 display irregularly-shaped clusters "a"-"k" simulated by Duczmal et al. [26]. Figure 3 illustrates the regions associated with six different irregularly shaped clusters for the new irural, imixed, and iurban benchmark data. Plots of the complete set of regions associated with the 45 irregularly-shaped clustering models are provided in the Supplementary Information.

The null model for the data (conditional on the total number of cases), is the same for every cluster model. In order to have a consistent null distribution across the benchmark data, Kulldorff et al. [25] simulated 99,999 data sets under the constant risk hypothesis, which assumes the risk of disease is the same in all regions. The total number of cases observed across the entire study area, y_A, was fixed at either 600 or 6000 cases for each simulation, depending on the cluster model. The simulated null data sets were simulated according to a multinomial distribution with y_A trials (cases) and the probability of observing a case in region i being n_i/n_A.

The relative risk (RR) of each cluster was chosen so that the null hypothesis (of no cluster) would be rejected with probability 0.999 using a standard binomial test assuming the cluster regions were known ahead of time. The mean associated with each cluster under the null and alternative hypotheses, as well as the associated relative risk under the alternative hypothesis, are provided in Tables 1 and 2.

Table 1: Properties of the benchmark data sets under the null and alternative hypotheses for circular clusters. The number of cases in the cluster is denoted by y_{in}.

			600 cases			6000 cases		
Model	Regions	Population	$E(y_{in}\|H_0)$	$E(y_{in}\|H_a)$	RR	$E(y_{in}\|H_0)$	$E(y_{in}\|H_a)$	RR
mixed	1	710196	14.43	39.33	2.85	144.27	207.53	1.45
	2	817050	16.60	42.66	2.69	165.98	233.14	1.42
	4	1108440	22.52	51.35	2.40	225.18	301.70	1.36
	8	1352284	27.47	58.32	2.24	274.71	358.05	1.32
	16	1684327	34.22	67.47	2.10	342.17	433.70	1.29
rural	1	2675	0.05	10.24	191.75	0.54	12.80	23.60
	2	22911	0.47	12.32	27.01	4.65	22.99	4.95
	4	132343	2.69	18.45	7.05	26.89	59.02	2.21
	8	204829	4.16	21.61	5.35	41.61	79.56	1.92
	16	360275	7.32	27.60	3.90	73.19	120.80	1.66
urban	1	786178	15.97	41.71	2.73	159.71	225.77	1.43
	2	1072181	21.78	50.29	2.43	217.81	293.25	1.36
	4	2953077	59.99	100.32	1.81	599.91	715.71	1.22
	8	5018909	101.96	149.86	1.63	1019.58	1162.02	1.17
	16	7627173	154.94	208.63	1.53	1549.44	1713.33	1.15

3.2 Implementation

The text files associated with the 81 cluster models created by Kulldorff et al. [25] and Duczmal et al. [26] were downloaded from the webpages indicated in Section 3.1, imported into R, and converted to a matrix. The regions included in each cluster were determined by considering the estimated risk for each region across all 10,000 simulated data sets. The cluster regions associated with each model were then verified numerically and graphically. The relevant null and cluster data sets are included in the **neastbenchmark** R package, which can be installed from `https://github.com/jfrench/neastbenchmark`. The **neastbenchmark** package [35] also includes several utility functions to automate the benchmarking process.

A spatial adjacency matrix must be specified for many of the methods previously discussed. Neither Kulldorff et al. [25] nor Duczmal et al. [26] provided a spatial adjacency matrix. Consequently, the **spdep** R package [36] was used to automatically construct a spatial adjacency matrix for the regions, where two regions were said to be connected if they shared a border. Additionally, manual connections were added between islands (e.g., Nantucket and Manhattan) and other regions based on seasonal ferry traffic or public transportation between the island and other regions. These traffic patterns may change over time, but this ensured that

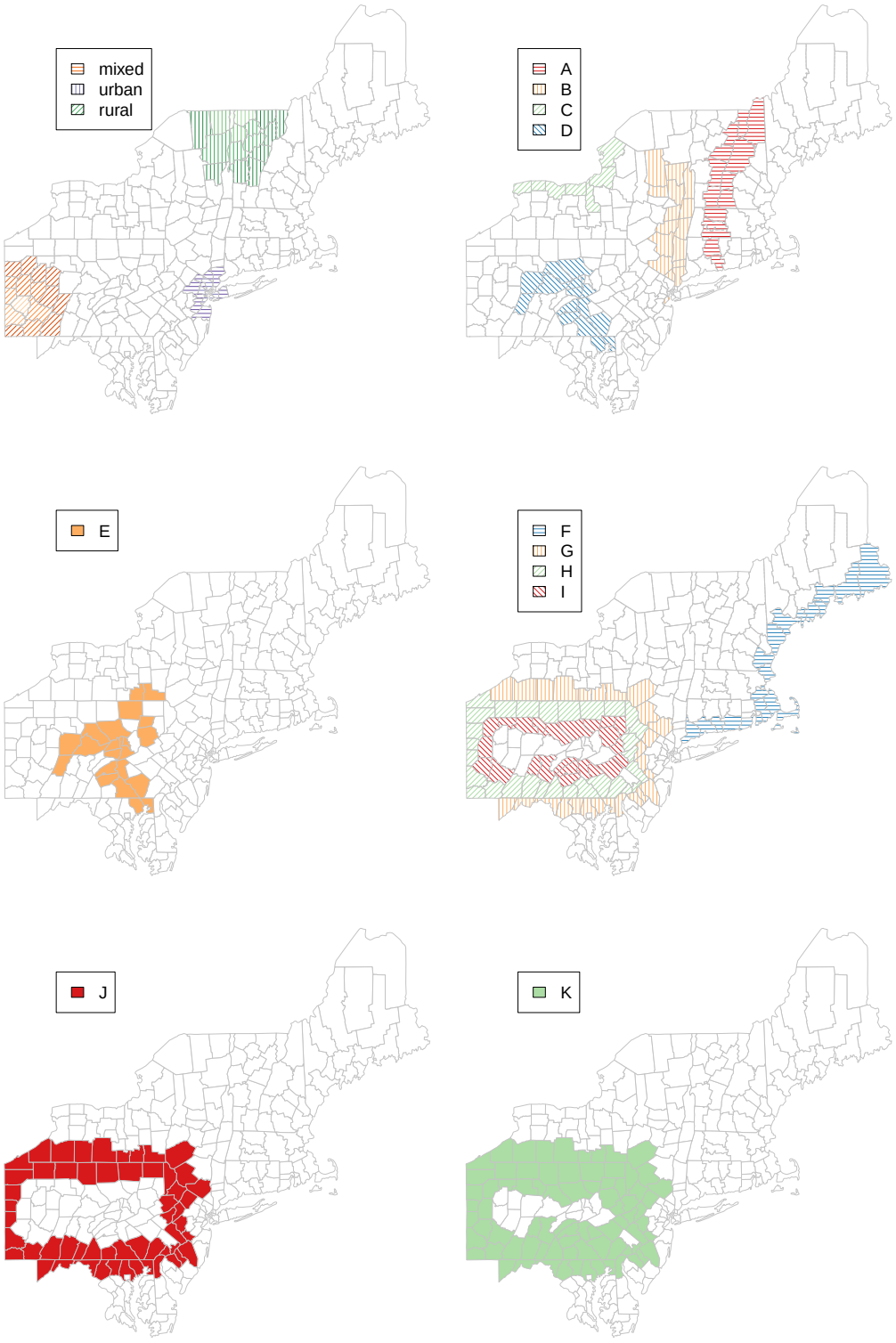

Figure 2: Maps indicating the different regions associated with the various cluster models. The differing shades of color for the mixed, urban, and rural regions are intended to help delineate the clusters comprised of 1, 2, 4, 8, or 16 regions.

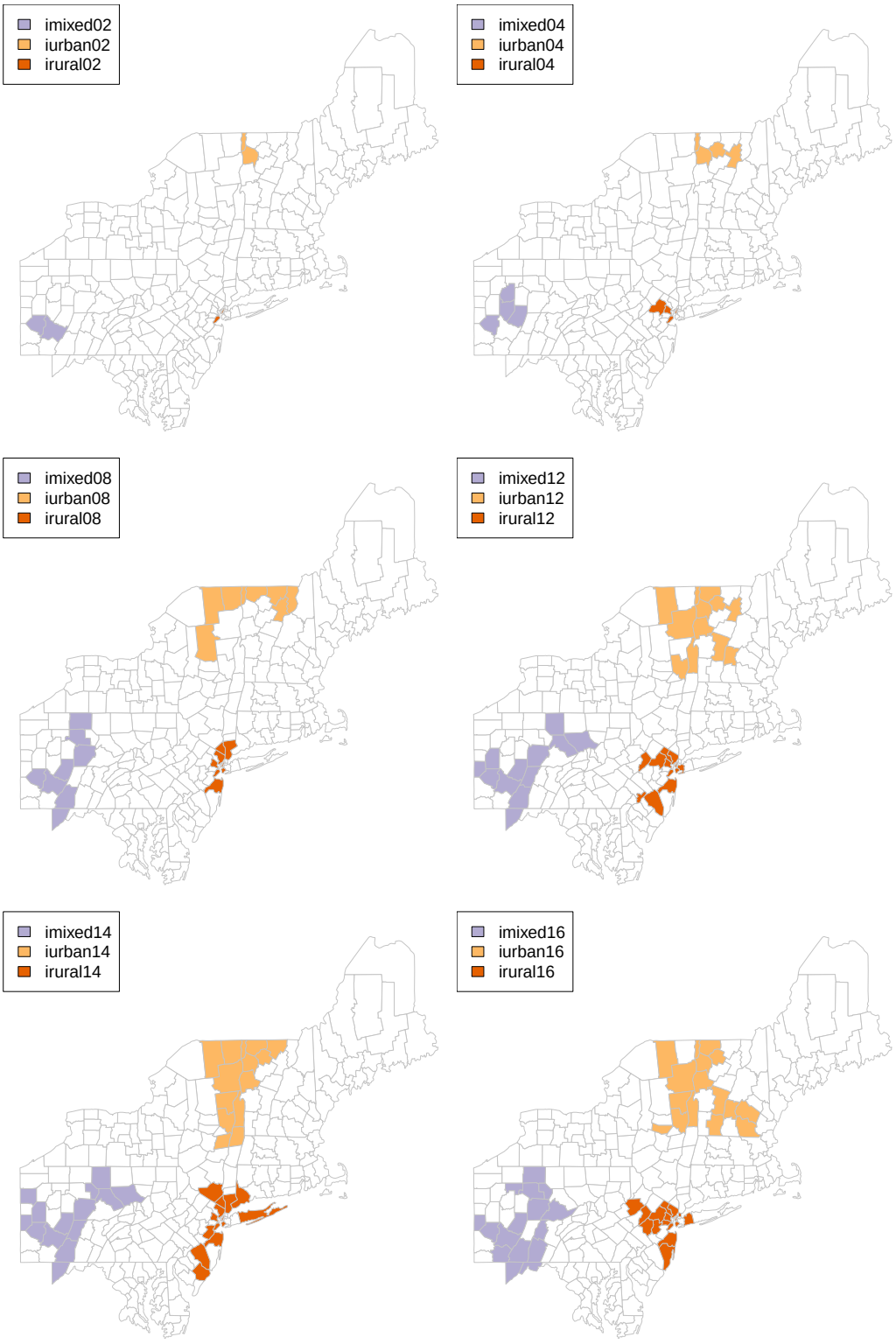

Figure 3: Maps indicating the different regions associated with the cluster models imixed, iurban, and irural comprised of 2, 4, 8, 12, 14, or 16 regions. Maps of all 45 clustering models can be found in Supplementary Information.

Table 2: Properties of the benchmark data sets under the null and alternative hypotheses for irregularly shaped clusters. The number of cases in the cluster is denoted by y_{in}.

		600 cases						600 cases			
Model	Regions	Population	$E(y_{in}\|H_0)$	$E(y_{in}\|H_a)$	RR	Model	Regions	Population	$E(y_{in}\|H_0)$	$E(y_{in}\|H_a)$	RR
imixed	2	903972	18.36	45.30	2.59	iurban	2	983454	19.98	47.68	2.51
	3	795354	16.16	41.99	2.72		3	2460200	49.98	87.87	1.89
	4	816964	16.60	42.66	2.69		4	1613284	32.77	65.54	2.12
	5	1075997	21.86	50.41	2.43		5	2842777	57.75	97.56	1.82
	6	1046467	21.26	49.54	2.45		6	3535465	71.82	114.66	1.74
	7	1129617	22.95	51.97	2.38		7	4989864	101.37	149.18	1.63
	8	1088254	22.11	50.76	2.42		8	3943710	80.12	124.52	1.70
	9	1192320	24.22	53.77	2.34		9	3957900	80.40	124.86	1.70
	10	1266637	25.73	55.90	2.29		10	4572707	92.89	139.44	1.65
	11	1249970	25.39	55.42	2.30		11	4725740	96.00	143.02	1.64
	12	1397049	28.38	59.57	2.22		12	5919393	120.25	170.52	1.58
	13	1426188	28.97	60.38	2.21		13	5343254	108.55	157.35	1.61
	14	1462539	29.71	61.39	2.19		14	5861993	119.08	169.22	1.59
	15	1435427	29.16	60.64	2.20		15	5642446	114.62	164.21	1.60
	16	1461010	29.68	61.35	2.19		16	5878543	119.42	169.59	1.59
irural	2	70417	1.43	15.34	10.98	a	14	1057407	21.48	49.86	2.44
	3	112107	2.28	17.49	7.88	b	16	1672387	33.97	67.15	2.10
	4	94514	1.92	16.62	8.87	c	7	709519	14.41	39.31	2.85
	5	196116	3.98	21.25	5.49	d	15	1119235	22.74	51.66	2.39
	6	169279	3.44	20.10	6.01	e	21	1483995	30.15	61.99	2.18
	7	164819	3.35	19.91	6.12	f	23	3198049	64.97	106.40	1.78
	8	119472	2.43	17.84	7.55	g	26	2477365	50.33	88.31	1.89
	9	238759	4.85	22.99	4.89	h	29	3112721	63.23	104.29	1.79
	10	248617	5.05	23.38	4.78	i	23	2185043	44.39	80.77	1.95
	11	322359	6.55	26.20	4.14	j	55	5590086	113.56	163.02	1.60
	12	339849	6.90	26.85	4.02	k	78	7775129	157.95	211.87	1.53
	13	367925	7.47	27.88	3.86						
	14	595522	12.10	35.66	3.07						
	15	437551	8.89	30.34	3.54						
	16	637030	12.94	37.00	2.98						

no region was isolated from all other regions. A connectivity map of the regions is shown in Figure 4. The associated connectivity matrix is `neastw` in the **neastbenchmark** R package.

The methods implemented in this paper are available in the **smerc** R package [24] (version 1.0 or higher). When possible (e.g., the circular, elliptic, flexible, and restricted flexible scan methods) the R implementations were compared with author-constructed reference implementations [17, 37] to ensure correctness of the implemented algorithms. Per the authors of the other benchmarked methods (dMST, double connection, m-link, ULS, fast subset), no available reference implementations currently exist.

All benchmark results utilized the Poisson version of the likelihood ratio statistic in Equation (4).

3.3 Power-related measures

There are numerous power-related measures that can be used in a cluster detection context. We discuss several that are relevant for the 10,000 benchmark data sets associated with each cluster model.

The most basic measure of power estimates the probability of rejecting H_0 when H_a is true, i.e., the probability of detecting a cluster when a cluster is present. Thus,

$$\text{power} = \frac{1}{10{,}000}\sum_{i=1}^{10{,}000} I(\text{a significant result was detected in benchmark data set } i),$$

where $I(\cdot)$ is the indicator function and a significant result occurs for benchmark data set i when the associated Monte Carlo p-value is less than a pre-specified threshold level α. The weakness of this measure is that it does not quantify whether the detected cluster overlaps the true cluster. The detected cluster could have no overlap with the true cluster and the basic power would be unaffected.

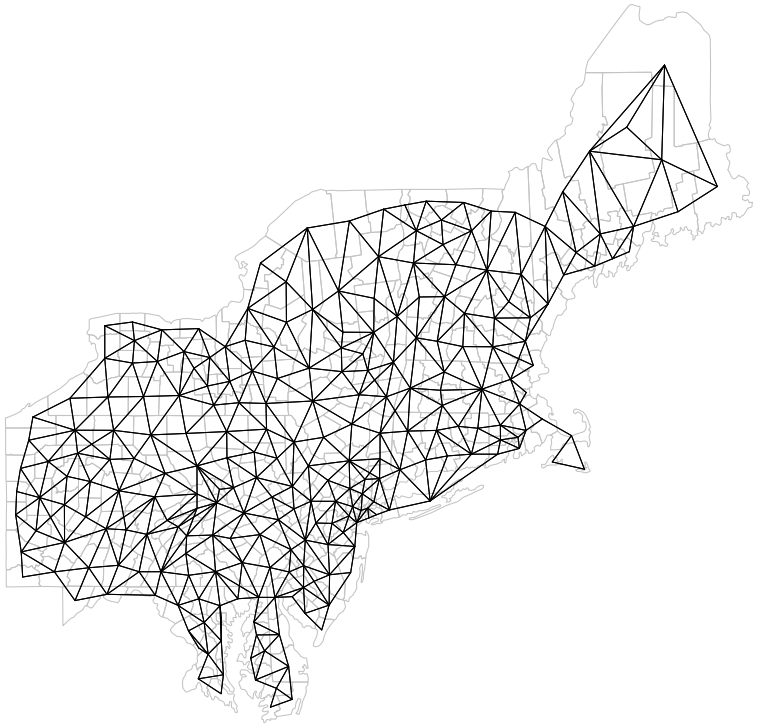

Figure 4: A connectivity map for the northeastern United States benchmark regions.

Another power-related measure is sensitivity. Let $\hat{Z}_i$ denote the detected cluster for benchmark data set i and C denote the actual cluster. Define $n(X)$ to be the population size of a zone X. Furthermore, define $p^{(i)}$ to be the Monte Carlo p-value for the largest test statistic (cf. Equations (5) and (6)) associated with benchmark data set i . The sensitivity is defined as the proportion of the true cluster captured by the detected cluster, averaged over all 10,000 simulations. More formally,

$$\text{sensitivity} = \frac{1}{10,000} \sum_{i=1}^{10,000} \frac{n(\hat{Z}_i \cap C)}{n(C)} I(p^{(i)} < \alpha). \tag{10}$$

The weakness of this measure is that sensitivity tends to increase with the size of the detected cluster, so methods that suffer from the octopus effect will have high sensitivity. We note that the sensitivity of a test for which a cluster is not detected is 0 since none of the true cluster was detected. Costa et al. [30] computed sensitivity (and other related measures) for a subset of the benchmark data for several tests, but did not consider whether $\hat{Z}_i$ was a significant cluster. Their calculation of sensitivity did not include the indicator function in Equation (10).

The positive predictive value (PPV) captures the proportion of the detected cluster that overlaps with the true cluster, averaged over all 10,000 simulations. Thus,

$$\text{PPV} = \frac{1}{10,000} \sum_{i=1}^{10,000} \frac{n(\hat{Z}_i \cap C)}{n(\hat{Z}_i)} I(p^{(i)} < \alpha).$$

A potential weakness of this measure is that a method with high PPV might detect only a small portion of the true cluster.

Lastly, we define the accuracy of a method as the proportion of the total population correctly categorized

by the method, averaged over all 10,000 simulations. The accuracy can be computed as

$$\text{accuracy} = \frac{1}{10{,}000}\sum_{i=1}^{10{,}000}\left(\frac{n(\hat{Z}_i \cap C) + n(\hat{Z}_i^c \cap C^c)}{n_A} I(p^{(i)} < \alpha) + \frac{n(C^c)}{n_A} I(p^{(i)} \geq \alpha)\right).$$

Note that the accuracy requires two cases depending on whether a cluster is detected. If a cluster is not detected, then all of the non-cluster population ($n(C^c)$) is detected accurately.

3.4 Benchmark results

We now discuss the results of the benchmark tests in terms of the power-related measures.

All tests were conducted using significance levels of $\alpha = 0.05$ and 0.01. Since neither significance level produced qualitatively different results, the results for $\alpha = 0.05$ are discussed below, while the results for $\alpha = 0.01$ are provided in Supplementary Information. No detected cluster was allowed to contain more than 50% of the at-risk population. The elliptic tests used the default values in SatScan [17], which considers shapes $\varsigma = 1, 1.5, 2, 3, 4$, and 5. The number of angles associated with each shape was 1, 4, 6, 9, 12, and 15, respectively. Elliptic tests were run with penalties of $\lambda = 0, 0.5$, and 1. To make computation feasible, the flexibly-shaped scan method was only run with $k = 15$ nearest neighbors. For the restricted flexible scan method, the middle p-value filtering level was set to $\alpha_1 = 0.2$. Additionally, to further facilitate computation, the restricted flexible scan method runs with $k = 90$. Since the largest clustering model includes 78 regions, $k = 90$ is reasonably large.

In the plots below, method labels are obvious except that elliptic0 refers to the elliptic test with penalty $\lambda = 0$, ellipticH refers to the elliptic test with penalty $\lambda = 0.5$, and elliptic1 refers to the elliptic test with penalty $\lambda = 1$. The label dc refers to the double connection test. All plots below use a significance level of $\alpha = 0.05$ when assessing the significance of a detected cluster.

Figure 5 displays violin plots of the average power for each method across all 126 cluster models, along with the jittered power value associated with each model. The circular, elliptic, flexible, double connection, and m-link methods all have similar power, with the restricted flexible scan method having slightly lower power. The dMST, edMST, fast subset, and ULS methods all have noticeably lower power.

Figure 6 displays violin plots of the average sensitivity for each method across all 126 cluster models, along with the jittered sensitivity value associated with each model. The sensitivity values vary considerably depending on the cluster model, with no clear winners. The circular, double connection, elliptic ($\lambda = 0$, 0.5, or 1), flexiblly-shaped, m-link, and restricted flexible methods all have similar sensitivities. The dMST, edMST, fast subset, and ULS methods have somewhat lower sensitivity.

The patterns for PPV and accuracy are qualitatively similar to those in Figure 5. These plots are provided in the Supplementary Information.

Overall, the circular, double connection, elliptic, flexible, m-link, and restricted flexible methods appear to perform better than their competitors, but it is difficult from Figures 5 and 6 to assess exactly how well they compare to each other. To better assess the differences, Figure 7 displays density plots of the average of the power-related measures for 6 of the methods (elliptic with $\lambda = 1$ was chosen to represent the elliptic scan method). From this figure, the circular and elliptic scan methods seem to have relatively the best overall performance. However, it must be noted that 70 of the 126 cluster models were designed for compatibility with the circular scan method (which is a special case of the elliptic scan method). Arguably, due to the high proportion of circular cluster models in the available benchmark data, those methods have the scales tipped in their favor.

To help mitigate this fact, we computed the power-related measures for the irregularly-shaped clusters alone (cluster models irural, imixed, iurban, a, b, …, k), then averaged the values for measure for each of the 56 cluster models. Similar to before, the circular, double connection, elliptic, flexible, m-link, and restricted flexible methods appear to perform better than their competitors. We display the results for the better performing methods in Figure 8, which displays boxplots of the 56 average values overlaid with a jitter plot of the actual values. Based on the available information in Figure 8, it appears that the circular, elliptic and m-link methods may have better sensitivity. On the other hand, the double connection and flexible scan methods have relatively better PPV. Additionally, all methods seem to yield similar results in terms of accuracy. We do not consider the average power to be a useful metric of method performance as it does not measure in any way how accurately the true cluster is detected.

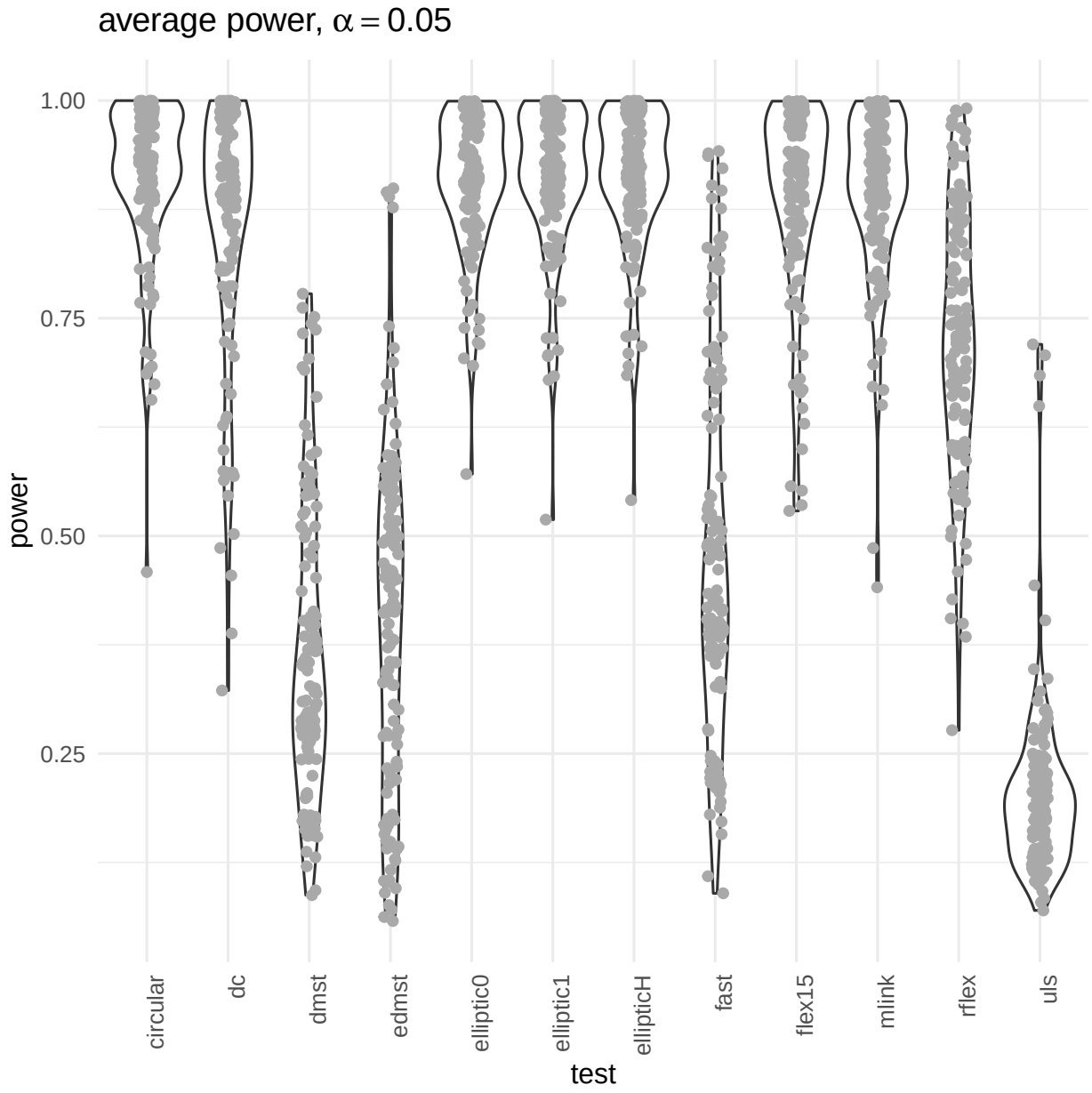

Figure 5: Average power of each method across all 126 cluster models.

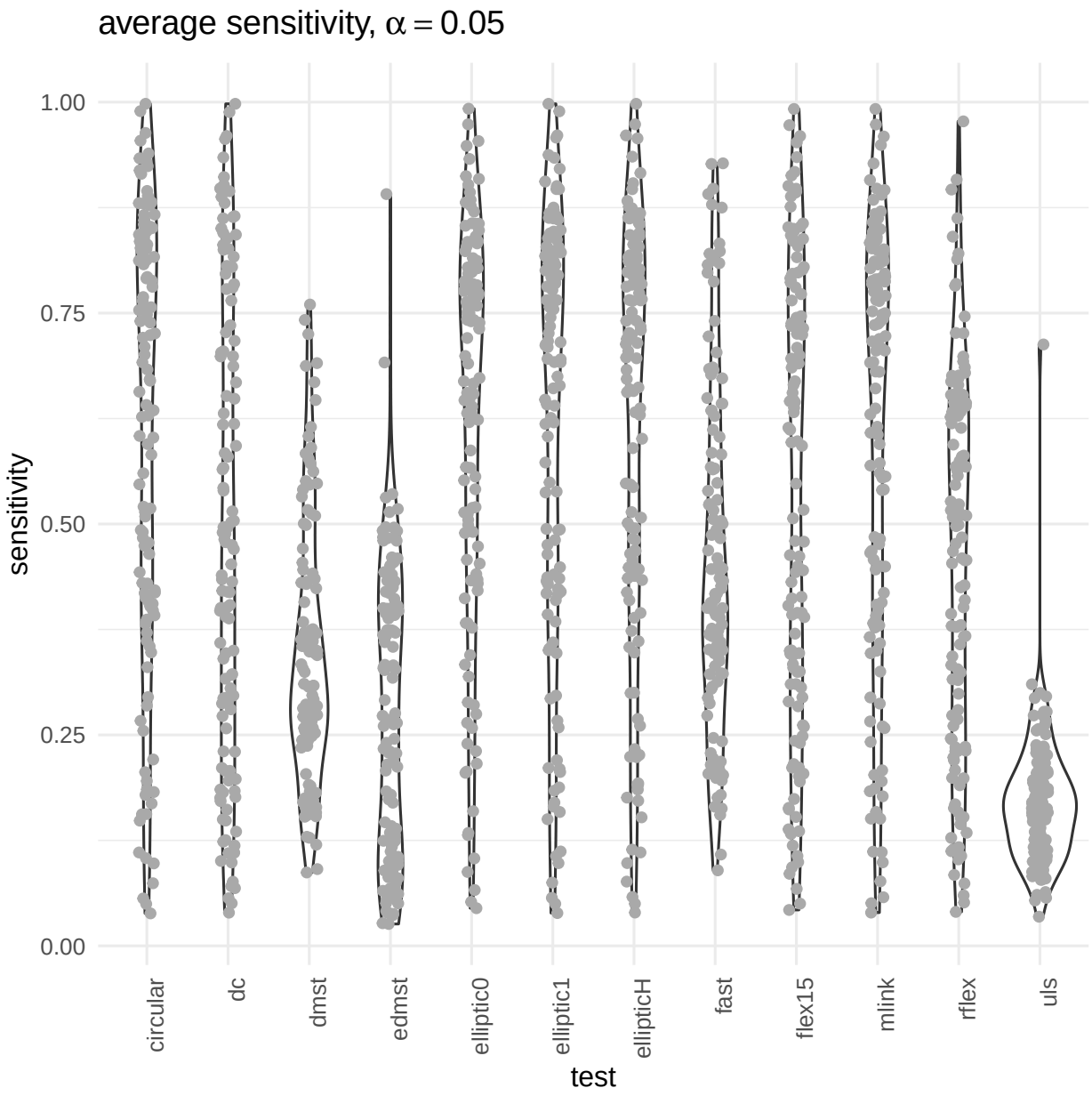

Figure 6: Average sensitivity of each method across all 126 cluster models.

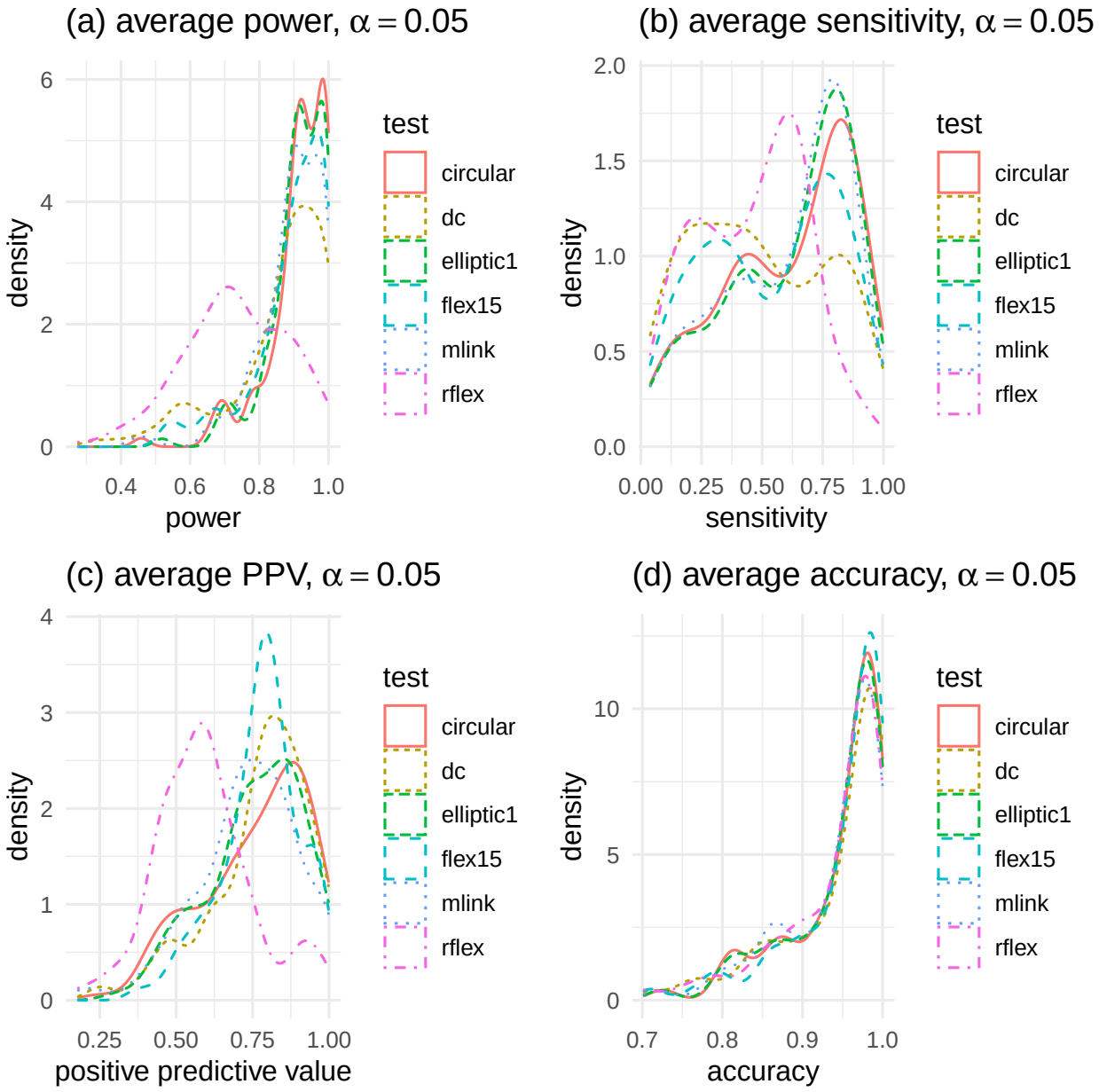

Figure 7: Power-related measures for several methods averaged across all 126 cluster models.

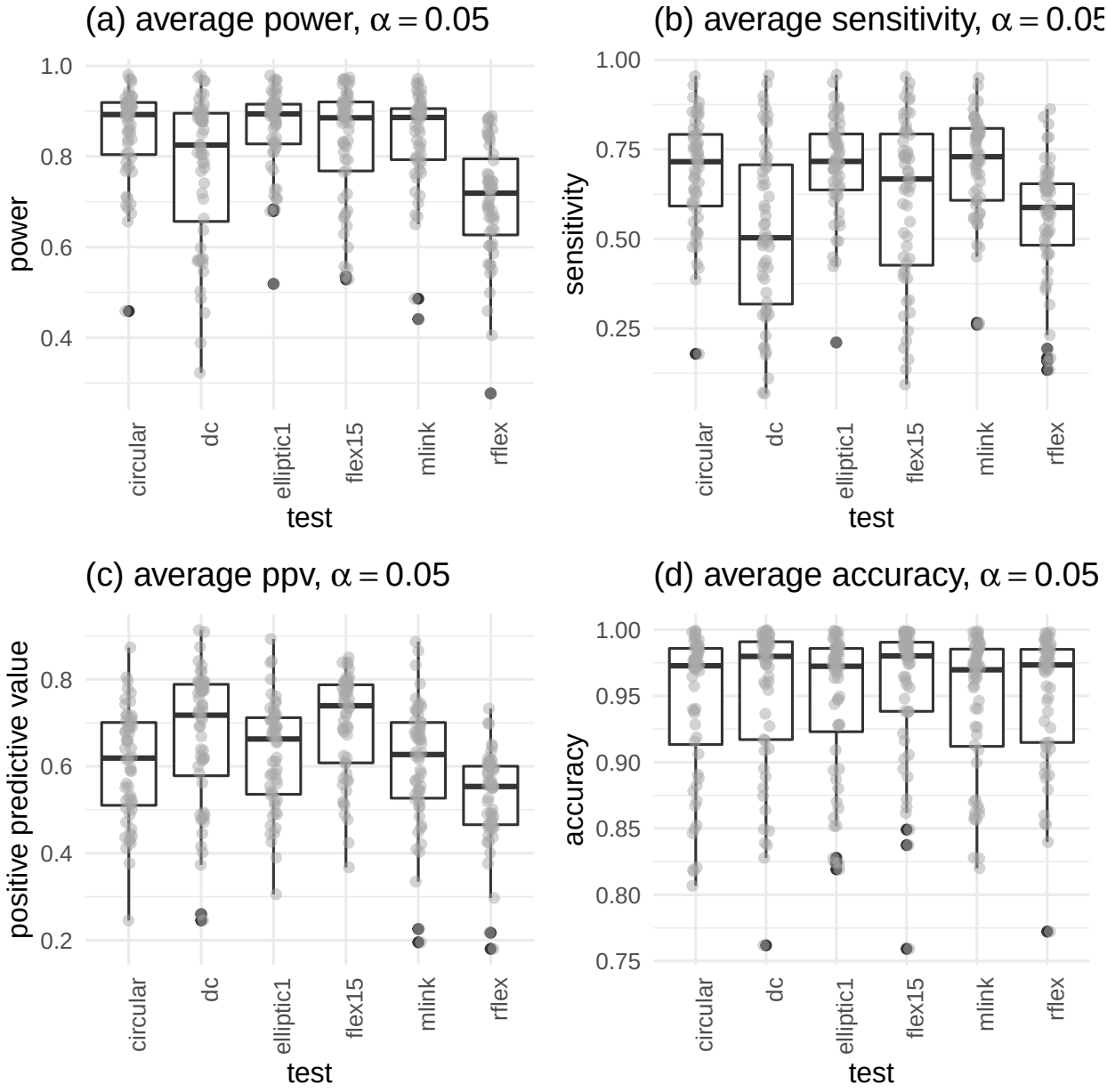

Figure 8: Power-related measures for several methods for the irregularly-shaped cluster models.

3.5 Benchmark timing

Computational efficiency can be an important consideration when deciding which scan method to apply to a data set. Each of the methods was applied to the original northeastern United States breast cancer data 10 times, with 99 data sets simulated under the null hypothesis. Table 3 displays the relative timings of each method in comparison to the smallest timing for the minimum execution time, the 0.25 quantile of the execution times, the average execution time, the median execution time, the 0.75 quantile of the execution times, and the maximum execution time. The fast subset scan is easily the fastest method. The circular scan

Table 3: Relative execution time statistics of each method when applied to the northeastern United States cancer data 10 times using 99 simulated null data sets. lq stands for the 0.25 quantile of the timings, while uq stands for the 0.75 quantile of the timings.

method	min	lq	mean	median	uq	max
fast	1.0	1.0	1.0	1.0	1.0	1.0
circular	30.7	31.0	31.2	31.1	32.0	30.8
uls	146.4	148.5	148.8	148.1	151.5	148.3
rflex	188.9	217.8	283.2	268.7	351.3	381.2
dc	329.5	332.0	332.8	328.4	330.6	353.3
edmst	1090.7	1093.5	1106.3	1113.7	1119.0	1099.8
elliptic0	1541.4	1589.1	1585.3	1574.0	1619.2	1618.4
elliptic1	1799.6	1869.5	1876.8	1906.2	1908.7	1833.4
ellipticH	2130.7	2136.6	2168.4	2163.9	2210.3	2160.2
mlink	18181.0	18482.2	18226.1	18230.3	18245.1	17738.2
flex15	18423.4	18553.0	18289.0	18278.4	18285.1	17545.5
dmst	26560.8	26739.0	26964.4	27304.3	27294.6	26071.0

method, the second fastest method, is approximately 30 times slower. The restricted flexible scan method is approximately 190 times slower than the fast subset scan method, while the double connection scan method is approximately 330 times slower. The maximum linkage and flexible scan method (with $k = 15$) take nearly 18,000 times as long as the fast subset scan method, while the dynamic minimum spanning tree method takes over 26,000 times as long. On a Windows 10 laptop with an Intel i7-7500U CPU running at 2.70GHz and 16 GB of RAM, the fast subset scan method results were returned instantaneously, the circular scan executed in less than a second, the elliptic scan method (any variant) took slightly under a minute to execute, while the dynamic minimum spanning tree method took a little over 7 minutes to execute.

We note that the timings above could change dramatically depending on the number of null data sets simulated. e.g., the restricted flexible scan method is typically quite fast, but an unusual “null" data set can increase its execution time by orders of magnitude. Additionally, methods that use a fixed set of candidate zones (specifically the circular, elliptic, and flexibly-shaped scan methods) scale better with a larger number of simulated null data sets because the construction of the candidate zones is a fixed cost independent of the number of simulated data sets. The execution time of the other methods is greatly impacted by the number of simulated null data sets because the candidate zones must be constructed independently for each simulated data set.

4 Discussion

We have performed extensive benchmark analysis on 10 different scan-based cluster detection methods: the circular, elliptic, flexibly-shaped, restricted flexible, dynamic minimum spanning tree, early-stopping dynamic minimum spanning tree, double connection, maximum linkage, fast subset, and upper level set. Each method was applied to 10,000 benchmark data sets from 126 different cluster models.

The circular, elliptic, flexible, restricted flexible, double connection, and maximum linkage methods performed noticeably better than the competing methods. No one method performed universally better than the other methods, but these methods distinguished themselves among the competitors.

The original circular scan method has relatively the best overall performance among the considered methods. It is very fast to execute and performed well in all performance-related measures. When considering only the irregularly-shaped clusters, the circular method still performed well. The main weakness of the circular scan method is that it may perform very poorly if the true cluster has a highly non-circular shape.

When looking at the performance of each method for both circular and irregularly-shaped clusters, the elliptic scan method seems to strike the best balance between sensitivity (the detected proportion of the true cluster) and positive predictive value (the proportion of the detected cluster that is part of the true cluster). Additionally, because the elliptic method has a fixed set of scanning windows, the elliptic scan method tends to scale well. The value chosen for the penalty parameter λ did not seem to have a substantial impact on performance. While the elliptic scan method has the potential to perform better than the circular scan method for irregularly-shaped clusters, it may not perform well if the true cluster has a highly non-elliptical shape.

In comparison to the elliptic method, the flexible and the double connection method provide better PPV. However, using only $k = 15$ regions for the flexible method may lead to a higher PPV while it might be reduced by larger k values. Additionally, the set of candidate zones for the double connection method is structured in such a way that more compact clusters are detected, which may result in higher PPV. However, when the elliptic, double connection, and flexible methods are compared both for PPV and sensitivity, a better overall performance is obtained by the elliptic scan method. Generally speaking, the double connection scan method performs similarly to the flexible scan method, and is arguably better in overall performance. We note that, in practice, it often executes more quickly than the flexible and restricted flexible scan method, especially if the restricted flexible scan method encounters a troublesome null data set. A weakness of the double connection method is that the double connection constraint has no mathematical basis. Philosophically, it makes sense that regions in the true cluster would be highly connected (i.e., share multiple borders) with other regions in the cluster, but there is no mathematical reason for this to occur. The double connection scan method can be expected to perform poorly when many regions in the true cluster have only a single connection to other regions in the true cluster.

The restricted flexible scan method had high overall accuracy for both the complete set of benchmark data and when focusing on the irregularly-shaped clusters. It also has a fairly fast execution time (in general). Its greatest strength is also its greatest weakness. By pre-filtering regions from the potential candidate zones with the α_1 tuning parameter, the method is able to improve the execution time by reducing the search space of the candidate zones while still retaining high performance. However, if certain connecting regions in a true cluster are pre-filtered, the method will perform poorly because those regions will not be considered in any candidate zone.

The effectiveness of the flexibly-shaped scan method is difficult to evaluate because of its high execution time. The method was evaluated with candidate zones containing no more than 15 total regions, which is much smaller than the true number of regions in many of the cluster models. In light of that, the method performed quite well. FlexScan [37], the reference implementation of the flexibly-shaped scan method, is substantially faster than the R implementation used in our analysis. However, that implementation was not suitable for use in this benchmark study because of the required user intervention and the output format of the results. The **rflexscan** [38] package implements the FlexScan algorithm in C++ with a convenient wrapper for data analysis in R. For a single data set, it is likely the flexibly-shaped scan method can be applied in a suitable amount of time. However, the execution time of the algorithm increases exponentially with the number of nearest neighbors. For $k = 15$ nearest neighbors, there are 1,158,378 candidate zones for the benchmark data. Increasing the number of nearest neighbors to $k = 20$ results in 26,665,183 candidate zones, more than a 20-fold increase in the number of candidate zones for only a small increase in the size of the candidate zones. Increasing the value of k for the flexibly-shaped scan method would provide it with a accurate measure of performance compared to competing methods.

The maximum linkage method had a strong overall performance. Similar to the double connection method, the maximum linkage connectivity constraint has no mathematical basis. Additionally, it takes substantially longer to execute than the other high-performing methods.

We generated 45 additional irregularly-shaped benchmark data sets irural, imixed, and iurban to make a fairer comparison and avoid biasing the results in favor of the circular scan method. Our overall conclusions are that the circular, elliptic, flexible, restricted flexible, double connection, and maximum linkage methods are strong choices for cluster detection. Due to their speed and performance, the circular and elliptic methods can be excellent first choices. They also have fast reference implementations available in SatScan [17]. If

adequate time is available, it may be wise to apply all six of these methods and compare their results to get a “preponderance of evidence” in assessing where a cluster may be located. In theory, one could combine the candidate zones from all methods simultaneously to get a “super test" for identifying the most likely cluster. However, this would dramatically increase execution time and it is not clear that the performance would improve (and may in fact be worse).

There are ways that this benchmark study could be improved, time allowing. Firstly, some of the tuning parameters were chosen to be smaller than desired in order for the benchmarking to be done in a suitable amount of time. A future study could increase these values. e.g., the number of nearest neighbors considered by the flexibly-shaped scan method should be increased, and the filtering parameter α_1 for the restricted flexible scan method should be increased.

All methods considered in this paper are implemented in the **smerc** R package [24], and the benchmark data are available in the **neastbenchmark** R package [35]. Script files to produce the simulation results are supplied as Supporting Information. The complete benchmark results are available as Supporting Information in a series of tables.

Acknowledgments

J. French was partially supported by NSF grants 1463642 and 1915277.

Author contributions

J. French contributed to the writing and revising of this manuscript, benchmark simulations, and developed the **smerc** and **neastbenchmark** packages [24, 35]. M. Meysami, L. Hall, and N. Weaver contributed to the writing and revising of this manuscript, as well as benchmark simulations. M. Nguyen and L. Panter contributed to the writing of this manuscript.

Financial disclosure

None reported.

Conflict of interest

The authors declare no potential conflict of interests.

Data availability

The data that support the findings of this study are openly available in the **neastbenchmark** R package [35].

Supporting information

The following supporting information is available as part of the online article:

1. R script files to produce the benchmark results and plots.
2. The complete benchmark results are available in a series of tables.

References

[1] Martin Kulldorff. A spatial scan statistic. *Communications in Statistics - Theory and Methods*, 26(6):1481–1496, 1997.

[2] Lance A Waller and Carol A Gotway. *Applied spatial statistics for public health data*. John Wiley & Sons, Hoboken, 2004.

[3] Toshiro Tango. *Statistical methods for disease clustering*. Springer-Verlag, New York, 2010.

[4] Roger S. Bivand, Edzer Pebesma, and Virgilio Gomez-Rubio. *Applied spatial data analysis with R, Second edition*. Springer, NY, 2013.

[5] Patrick AP Moran. Notes on continuous stochastic phenomena. *Biometrika*, 37(1/2):17–23, 1950.

[6] Robert C Geary. The contiguity ratio and statistical mapping. *The Incorporated Statistician*, 5(3):115–146, 1954.

[7] Neal Oden. Adjusting moran's i for population density. *Statistics in Medicine*, 14(1):17–26, 1995.

[8] Thomas Waldhör. The spatial autocorrelation coefficient moran's i under heteroscedasticity. *Statistics in Medicine*, 15(7-9):887–892, 1996.

[9] Renato M Assuncao and Edna A Reis. A new proposal to adjust moran's i for population density. *Statistics in medicine*, 18(16):2147–2162, 1999.

[10] Stan Openshaw, Martin Charlton, Alan William Craft, and JM Birch. Investigation of leukaemia clusters by use of a geographical analysis machine. *The Lancet*, 331(8580):272–273, 1988.

[11] Gerard Rushton and Panos Lolonis. Exploratory spatial analysis of birth defect rates in an urban population. *Statistics in Medicine*, 15(7-9):717–726, 1996.

[12] Eric J. Turnbull, Bruce W.and Iwano, William S. Burnett, Holly L. Howe, and Larry C. Clark. Monitoring for clusters of disease: Application to leukemia incidence in upstate new york. *American Journal of Epidemiology*, 132(supp1):136–143, 1990.

[13] Julian Besag and James Newell. The detection of clusters in rare diseases. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 154(1):143–155, 1991.

[14] Sara Mclafferty. Disease cluster detection methods: recent developments and public health implications. *Annals of GIS*, 21(2):127–133, 2015.

[15] Martin Kulldorff and Neville Nagarwalla. Spatial disease clusters: Detection and inference. *Statistics in Medicine*, 14(8):799–810, 1995.

[16] Google Scholar. `https://scholar.google.com`, 2021. Accessed November 12, 2021.

[17] Martin Kulldorff. SaTScan, version 9.6. `https://satscan.org`, 2018.

[18] Martin Kulldorff, Lan Huang, Linda Pickle, and Luiz Duczmal. An elliptic spatial scan statistic. *Statistics in Medicine*, 25(22):3929–3943, 2006.

[19] Ganapati P Patil and Charles Taillie. Upper level set scan statistic for detecting arbitrarily shaped hotspots. *Environmental and Ecological statistics*, 11(2):183–197, 2004.

[20] Toshiro Tango and Kunihiko Takahashi. A flexibly shaped spatial scan statistic for detecting clusters. *International journal of health geographics*, 4(1):11, 2005.

[21] Toshiro Tango and Kunihiko Takahashi. A flexible spatial scan statistic with a restricted likelihood ratio for detecting disease clusters. *Statistics in medicine*, 31(30):4207–4218, 2012.

[22] R. Assunção, M. Costa, A. Tavares, and S. Ferreira. Fast detection of arbitrarily shaped disease clusters. *Statistics in Medicine*, 25(5):723–742, 2006.

[23] Daniel B Neill. Fast subset scan for spatial pattern detection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(2):337–360, 2012.

[24] Joshua P. French. smerc: Statistical methods for regional counts. `https://cran.r-project.org/package=smerc`, 2021. R package version 1.4.

[25] Martin Kulldorff, Toshiro Tango, and Peter J Park. Power comparisons for disease clustering tests. *Computational Statistics & Data Analysis*, 42(4):665–684, 2003.

[26] Luiz Duczmal, Martin Kulldorff, and Lan Huang. Evaluation of spatial scan statistics for irregularly shaped clusters. *Journal of Computational and Graphical Statistics*, 15(2):428–442, 2006.

[27] Luiz Duczmal and Renato Assunção. A simulated annealing strategy for the detection of arbitrarily shaped spatial clusters. *Computational Statistics & Data Analysis*, 45(2):269–286, 2004.

[28] G. P. Patil, S. W. Joshi, and R. E. Koli. Pulse, progressive upper level set scan statistic for geospatial hotspot detection. *Environmental and Ecological Statistics*, 17(2):149–182, Jun 2010.

[29] Toshiro Tango. A spatial scan statistic with a restricted likelihood ratio. *Japanese Journal of Biometrics*, 29(2):75–95, 2008.

[30] Marcelo Azevedo Costa, Renato Martins Assunção, and Martin Kulldorff. Constrained spanning tree algorithms for irregularly-shaped spatial clustering. *Computational Statistics & Data Analysis*, 56(6):1771–1783, 2012.

[31] Luiz Duczmal, André L.F. Cançado, Ricardo H.C. Takahashi, and Lupércio F. Bessegato. A genetic algorithm for irregularly shaped spatial scan statistics. *Computational Statistics & Data Analysis*, 52(1):43 – 52, 2007.

[32] Luiz Duczmal, Andre L. F Cançado, and Ricardo H. C Takahashi. Delineation of irregularly shaped disease clusters through multiobjective optimization. *Journal of Computational and Graphical Statistics*, 17(1):243–262, 2008.

[33] Alan T. Murray, Anthony Grubesic, and Ran Wei. Spatially significant cluster detection. *Spatial Statistics*, 10:103–116, 11 2014.

[34] Martin Kulldorff, Eric J. Feuer, Barry A. Miller, and Laurence S. Freedma. Breast cancer clusters in the northeast united states: A geographic analysis. *American Journal of Epidemiology*, 146(2):161–170, 1997.

[35] Joshua P. French. neastbenchmark: Benchmark data for disease clusters. `https://github.com/jfrench/neastbenchmark`, 2021. R package version 0.2.

[36] Roger Bivand and Gianfranco Piras. Comparing implementations of estimation methods for spatial econometrics. *Journal of Statistical Software*, 63(18):1–36, 2015.

[37] Kunihiko Takahashi, Tetsuji Yokoyama, and Toshiro Tango. FleXScan user guide. `https://sites.google.com/site/flexscansoftware/`, 2010.

[38] Takahiro Otani and Kunihiko Takahashi. Flexible scan statistics for detecting spatial disease clusters: The rflexscan R package. *Journal of Statistical Software*, 99(13):1–29, 2021.